Reduced synonymous substitution rate at the start of enterobacterial genes

Adam Eyre-Walker and Michael Bulmer
Department of Biological Sciences, Rutgers University, Piscataway, NJ 08855-1059, USA

Received May 24, 1993; Revised and Accepted August 10, 1993

ABSTRACT

Synonymous codon usage is less biased at the start of *Escherichia coli* genes than elsewhere. The rate of synonymous substitution between *E.coli* and *Salmonella typhimurium* is substantially reduced near the start of the gene, which suggests the presence of an additional selection pressure which competes with the selection for codons which are most rapidly translated. Possible competing sources of selection are the presence of secondary ribosome binding sites downstream from the start codon, the avoidance of mRNA secondary structure near the start of the gene and the use of sub-optimal codons to regulate gene expression. We provide evidence against the last of these possibilities. We also show that there is a decrease in the frequency of A, and an increase in the frequency of G along the *E.coli* genes at all three codon positions. We argue that these results are most consistent with selection to avoid mRNA secondary structure.

INTRODUCTION

Non-random usage of synonymous codons is well-documented. It is most pronounced in highly-expressed genes in unicellular organisms, and has been extensively studied in *Escherichia coli* and *Saccharomyces cerevisiae*[1,2]. Codons which are translated rapidly, because they are recognized by an abundant tRNA with a natural match in the wobble position of the anticodon, are used preferentially in highly-expressed genes, while codons recognized by a rare tRNA tend to be avoided[1,3]. It seems likely that the preferential use of rapidly translated codons is determined by selection upon the rate of elongation during translation[4,5], though the exact nature of the selection pressure is not fully understood[6,7].

In *E.coli*, though not in *Bacillus subtilis*[8] and *S.cerevisiae*, codon usage bias is weaker at the start of the gene than elsewhere[9,10]. A possible explanation for this observation is the presence of other selection pressures at the start of the gene which compete with selection for rapid elongation, such as the use of fast or slow codons to modulate gene expression[10,11], the presence of a secondary ribosome binding site[12–14] or constraints upon mRNA secondary structure[15–17]. In this paper we show that the silent substitution rate between *E.coli* and *Salmonella typhimurium* is reduced at the start of the gene, suggesting that there are additional selection pressures in this region, and we attempt to determine what these selection pressures might be.

METHODS

Data sets

Three data sets were used in this investigation. The first was a compilation of 1210 complete *E.coli* sequences from the Ecoseq6 database[18] (Dr Kenn Rudd pers comm). The second was a set of 1014 gene starts covering the region from 100 base pairs downstream of the first base in the gene, to 100 base pairs upstream, from Ecoseq5[18] (Dr Kenn Rudd pers comm). The third data set was a collection of 138 aligned *E.coli* and *S.typhimurium* genes extracted from Genbank release 74 using the GCG sequence analysis package[19]. Complete sequences were used where available, but for a small number of genes only the 5′ portion had been sequenced. All genes used were longer than 300 base pairs. Details of the *E.coli/S.typhimurium* datasets are available on request from the authors. All data sets were split into three groups according to their codon adaptation index (C-AI, see below), less than 0.33, greater than 0.33 but less than 0.41, and greater than 0.41. This split the data set of aligned sequences into three roughly equal sized groups: 41 genes in the low CAI group, 51 in the medium CAI group, and 46 in the high CAI group. The large *E.coli* data set was divided into groups of 447, 371 and 392 genes respectively. The data sets were so divided to study any possible effects of gene expression, gene expression being correlated to CAI value.

Codon bias and substitution rates

The codon adaptation index (CAI) was calculated by the method of Sharp and Li[20] with modifications as before[3]. The synonymous substitution rate was calculated as the proportion of third position sites in a gene which differed between the two species, in codons which showed no non-synonymous differences. In most calculations values were averaged over groups of 5 consecutive codons to reduce sampling error.

RESULTS

Codon usage bias and intragenic position

The codon adaptation index (CAI) is a measure of the degree of bias of synonymous codon usage in favor of optimal codons, having a maximal value of 1 when only optimal codons are used. Figure 1 shows the CAI as a function of codon position along the gene (after the start codon) for 1210 *E.coli* genes, divided into three groups by their overall CAI. These three groups, with high, medium and low CAI values may be taken to represent genes with high, medium and low expression levels. For example, the high CAI group includes the outer membrane protein *ompA* and ribosomal protein genes, while the low CAI group includes the regulatory genes such as *lexA* and *dnaG*. To reduce sampling noise codons have been combined in groups of five, 1–5, 6–10, and so on, codon 1 being the first codon after the start codon.

This Figure confirms previous results based on a much smaller data set[9]. There is a marked reduction in the CAI at the start of the gene, which is most marked in highly expressed genes with strong codon usage bias and is barely detectable in lowly expressed genes. The CAI reaches a plateau at about the 100th codon. (Data not shown after the 300 th codon do not show any change in the average CAI.)

In interpreting these results it is useful to know as a benchmark the CAI which would be expected under random usage of synonymous codons. To calculate this we have imagined a very long gene with the same amino acid usage as observed in the total sample of genes but with bases in silent positions determined randomly given the genomic GC composition of 51%. The random CAI calculated in this way is 0.240. Thus the genes in the high CAI group have a strong bias towards the usage of optimal codons except near their start (remembering that the maximum possible CAI is 1), while the genes in the low CAI group have only a slight tendency to use optimal codons. The reduction in the CAI near the start of the gene is proportional to the degree of non-randomness in the rest of the gene. This is consistent with either a reduction in selection upon elongation rate or with a competing selective force near the start of the gene. However, it does not fit well with the idea that codon usage near the start of the gene is used to modulate gene expression, since in that case one might expect to find CAI values below the random level in weakly expressed genes if non-optimal codons were being used to reduce expression level. Additional evidence against this hypothesis is presented below in the section on 'The similarity of synonymous substitutions'.

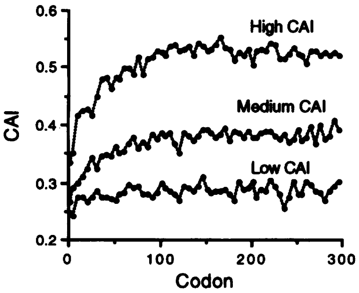

Figure 1. CAI value plotted against position for a set of *E.coli* genes split into three groups according to their CAI value. Each point represents the average of five codons. The start codon has been omitted.

Synonymous substitution rate and intragenic position

Figure 2 shows the synonymous substitution rate between *E.coli* and *S. typhimurium* (calculated as the uncorrected proportion of synonymous substitutions in codons with no non-synonymous change) in 138 genes sequenced in both species. The rates have been plotted by position after the start codon for three expression levels, assessed by the CAI; to reduce sampling noise positions have been combined in groups of five as before. The substitution rate is rather lower in highly expressed than in weakly expressed genes, which presumably reflects a stronger selective constraint on synonymous codon usage in highly expressed genes[21]. The substitution rate is substantially lower in all three expression levels near the start of the gene. The reduction in substitution rate is most pronounced in the first five codons (the first point plotted), but extends to a lesser degree through the first fifty codons. Note that although the trends in CAI and substitution rate appear to extend over different distances, the CAI calculations are probably subject to much less error since the data set is much larger.

The reduction in the synonymous substitution rate at the beginning of the gene suggests increased selection pressure there to maintain the *status quo*. This is inconsistent with the first explanation for the reduction in the CAI at the start of the gene, that there is a reduction in the selection pressure upon elongation rate. We conclude that there is an additional selection pressure on sites near the start of a gene which competes with selection

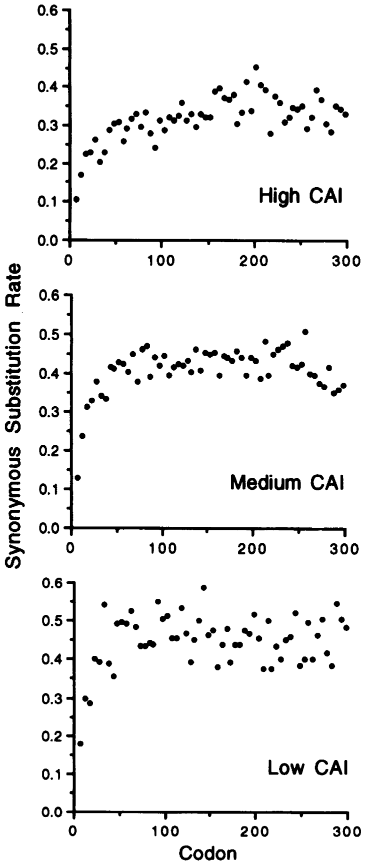

Figure 2. The synonymous substitution rate plotted against position for the three groups of genes. Each point represents the average of five codons.

for translationally optimal codons, thus reducing the CAI towards the level expected under random codon usage.

The similarity of synonymous substitutions

It is possible that genes use codons near the beginning to modulate gene expression by increasing or decreasing the rate of translation through them[10,11]. A newly arrived ribosome would therefore block the ribosome binding site for a shorter or longer period, which would in turn increase or decrease the expression level of the gene. This would explain the lower substitution rate near the start of the gene, but it does not provide a plausible explanation of the pattern of the relationship between the CAI and codon position. One would rather expect to find the CAI falling below the 'random' level in weakly expressed genes if they are using rare codons to decrease their expression[22].

This idea can also be tested by studying the nature of the substitutions near the start of the gene. Suppose that different codons occur at a particular position in *E.coli* and *S. typhimurium* respectively, and that the relative usages of these synonymous codons in a reference set of highly expressed genes are w_1 and w_2 respectively. The similarity of the substitution with respect to elongation rate can be measured by the index $|\log(w_1/w_2)|$; a value of 0 indicates that the two codons have the approximately the same elongation rate since they are used equally frequently in highly expressed genes, whereas a large value of the index indicates that the codons differ substantially in the speed with which they are translated. If there is selection to maintain a particular elongation rate (either for a high rate away from the start of the gene, or perhaps for a low rate near the start of the gene to reduce gene expression) then one would expect the substitutions that occur to be more similar than in the absence of such selection.

Table 1 shows the average value of this index for those codons which have undergone a synonymous substitution. The last column shows that, away from the start of the gene, the index decreases (i.e. the substitutions become more similar with regard to elongation rate) as the degree of gene expression increases, in accordance with the idea that selection for efficiently translated codons increases with the level of gene expression. We may take the value of the index in low CAI genes away from the start of the gene (1.83) as the neutral value to be expected in the absence of selection for translational efficiency, since there is nearly random codon usage in these genes. The first column shows that codons near the start of the gene have this neutral value of the index regardless of their CAI value, suggesting that selection upon elongation rate (either for or against it) is relatively weak in comparison with other selection pressures. There is thus no support for the idea that codons near the start of the gene are used to modulate the level of gene expression by promoting slow or fast translation through them.

Compositional changes

In order to understand further the causes of the low synonymous codon bias and substitution rate at the start of *E.coli* genes we calculated the composition of each base position. The results for the third codon position of the whole data set are shown in Figure 3. There are increases in G and C, and a decrease in A along the gene sequences, whereas T shows little or no change. Similar trends, in particular those for A and G are seen at both the first and second codon positions, although they are not as strong at the second. The trends are also seen in the high, medium and low CAI groups of genes and when four-fold degenerate codons are considered separately (note that since a certain fraction of third positions are two-fold degenerate, the base compositions of A and G, C and T are not independent). Stormo *et al.*[23] and Schneider *et al.*[24] have also looked at the base composition at the start of bacterial sequences. Although they never drew attention to the fact, the trends shown in figure 3 are evident in their data, though less clearly marked. This is not surprising since their data set was considerably smaller than the one used here, they used a mixture of bacterial and phage sequences and they did not consider the three codon positions separately. The three codon positions show similar trends in composition but over different absolute values; for instance the frequency of G from the 20th codon onwards is 35% at the first codon position, 18% at the second and 26% at the third; thus any trend will be obscured if all codon psoitions are considered simultaneously.

There are at least two potential explanations of these trends in composition. First there could be sequences downstream from the start codon which bind the ribosome. Sprengart *et al.*[13] have suggested that binding the 16S RNA sequence 3′-AGUACUUA-GUGUUUC-5′ could be important for gene expression. Since this sequence is U rich one would expect the frequency of A to be greater near the start of genes, as we observe. However if we note that G can bind both C and U, the increase in the frequency of G along the gene does not appear to be predicted by binding to the 'Sprengart' sequence. Similarly the the 16S RNA sequence suggested by Petersen *et al.*[12] 3′-AGUUUGAG-AAGUUAAA-5′ fails to explain all the trends in base composition. Since the sequence is A rich one would expect the

Table 1. Similarity of synonymous substitutions ± SE

	Codons 1–10	Codons 61–300
Low CAI genes	1.83 ± 0.16	1.83 ± 0.02
Medium CAI genes	1.83 ± 0.15	1.76 ± 0.02
High CAI genes	1.77 ± 0.16	1.54 ± 0.02
All genes	1.81 ± 0.09	1.72 ± 0.01

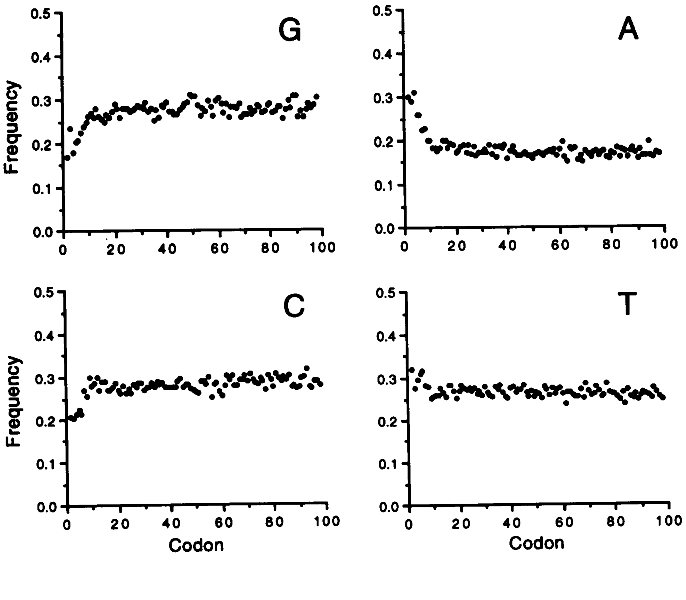

Figure 3. The third position base composition plotted against position. Each codon is plotted separately. The start codon has been omitted.

largest trend to be for T, but this is not seen. However the sequence contains no C so if G:U base pairs are weak one would expect the frequency of G to be low at the start of the gene, as is observed.

Second there may be selection against the formation of mRNA secondary structure at the start of the gene to allow ribosome binding[15-17]. Since G can bind C strongly and U weakly we would expect it to be the least favoured base in areas of low secondary structure potential, whereas A, which can only bind U weakly should be the most favoured base, as we see in the data. To test this idea further we calculated the correlation between G/(G+C) upstream from the start codon with G/(G+C) downstream. A similar calculation was performed for A/(A+T). Thirty base pairs either side of the start codon were used in each case since the main trend in composition is seen over the first 10 codons. If there is selection to avoid secondary structure around the start codon and Shine Dalgarno sequence then we would expect G/(G+C) upstream to be positively correlated to G/(G+C) downstream; for instance if the start codon is involved in a stem loop struture that must involve binding between the bases upstream with the bases downstream of the start codon. Although this is clearly a weak test, the correlations for both G/(G+C) and A/(A+T) turned out to be highly significant (Pearson's correlation coefficients are 0.140 and 0.105 respectively, $p<0.0001$). Thus there is some evidence that selection acts to minimise the amount of secondary structure at the start of the mRNA.

An intriguing question is why the trends in base composition do not extend as far as either the trends in substitution rate or CAI, especially when the analysis was performed on the larger data set. There are at least two possibilities. First, there may be more than one factor reducing synonymous codon bias at the start of enterobacterial genes. Second, the trends in composition may simply be a weak signal for the process that is responsible for reduction in codon bias. For instance let us imagine that selection is acting to minimise mRNA secondary structure at the start of the gene. Generally for any amino acid there is only one optimal codon, but to avoid a particular structure one might be able to use any one of two or three bases. For instance to avoid base pairing with A one can use A, G or C. Therefore we would expect the trend in base composition to be less obvious than that for codon bias.

SUMMARY AND CONCLUSIONS

We have confirmed, using a very large data base, the previous observation that *E.coli* genes tend to use translationally non-optimal codons more frequently near the start of the gene than elsewhere[9,10]. We have also found a substantial reduction in the synonymous substitution rate between *E.coli* and *S. typhimurium* near the start of the gene, suggesting an increased selection pressure at synonymous sites.

One possible explanation of these results is that there is selection for reduced elongation rate near the start of the gene to modulate the level of gene expression[10,11]. Three lines of evidence argue against this interpretation. First, the pattern of the relationship between the CAI and position in Figure 1 shows that the CAI (which measures the extent of the preference for optimal codons) falls at the start of the gene towards the level expected under random codon usage but never falls below it. Second, we have presented in Table 1 an index of the similarity of synonymous substitutions with regard to elongation rate. This index shows no evidence of conservation of elongation rate (either for fast or slow translation) near the start of the gene. Third, it is difficult to explain the changes in composition at the start of *E.coli* genes in terms of selection for slow elongation, since the trends in composition are seen independently of any change in CAI in the low CAI group of genes, and both the high and low CAI group of genes have similar compositions in the main body of the gene (results not shown).

Our conclusion that sub-optimal codons are not used at the start of a gene to regulate gene expression seems to conflict with the results of Chen and Inouye[10]. They showed that placing five translationally inefficient arginine codons near the start of the *lacZ* gene reduced gene expression substantially. Three points are pertinent. First, *E.coli* may not use a form of gene regulation even if it exists. Second, Chen and Inouye[10] only detected a substantial effect on gene expression when four or five rare codons were inserted next to each other; this is rarely found. And third, caution must be excercised in interpreting such studies if the possible effects on mRNA secondary structure cannot be controlled for, since even very small changes in the potential to form such structures can substantially reduce gene expression.[15,16]

We conclude that the reduced synonymous substitution rate near the start of bacterial genes reflects an additional selection pressure acting in this region which is uncorrelated with elongation rate. Selection for this factor competes with selection for elongation rate and therefore reduces the use of optimal codons towards the random level. We believe that this selective force is likely to act on mRNA function. It might, for example, be selection against the formation of secondary structure in the messenger[15-17], which would interfere with ribosome binding near the start of the gene, or it might reflect the presence of additional ribosome binding sites after the start codon[12-14]. Resolving these two possibilities is likely to be difficult[15].

Interestingly the present work suggests that sites important for initiation may lie further into the coding sequence than was previously thought. The ribosome appears to cover bases -20 to $+13$ in protection assays[25] which is the same range of positions over which non-random base compositions have been previously detected[23,24]. However we have shown that there are trends in composition out to at least the 30th base, an increase in sequence conservation out to the 30th codon (90th base) and trends in codon bias out to the 100th codon (300th base).

ACKNOWLEDGEMENTS

We are grateful to Kenn Rudd and Paul Sharp for their help with the compilation of the data sets; and to Masayori Inouye, Siv Andersson, Chuck Kurland and two anonymous referees for their comments on this manuscript.

REFERENCES

1. Ikemura,T. (1985). *Mol Biol Evol.*, **2**, 13–34.
2. Sharp,P. (1989). In Hill,W.G. and McKay,T. (ed.), Evolution and animal breeding: reviews on molecular and quantitative approaches in honour of Alan Robertson. Wallingford CAB International.
3. Bulmer,M. (1988). *J Evol Biol.*, **1**, 15–26.
4. Pedersen,S. (1984). *EMBO J.* **3**, 2895–2898.
5. Srensen,M.A., Kurland,C.G., and Pedersen,S. (1989). *J. Mol. Biol.* 207,365–377.
6. Bulmer,M. (1991). *Genetics.*, **129**, 897–907.
7. Kurland,C.G. (1991). *FEBS Letts* 285, 165–169.

8. Sharp, P.M., Higgins, D.G., Shields, D.C., Devine, K.M., and Hoch, J.A. (1990). In Zulowski, M.M., Ganesan, A.T., and Hoch, J.A. (ed.), Genetics and Biotechnology of Bacilli, vol. 3. Academic Press.
9. Bulmer,M. (1988). *J Theor Biol.*, **133**, 67–71.
10. Chen, G-F.T., and Inouye,M., (1990). *Nucleic Acid Res.* **18**, 1465–1473.
11. Konisberg,W.J.N., and Godson,G.N., (1983). *Proc Natl Acad Sci USA.*, **80**, 687–691.
12. Petersen,G.B., Stockwell,P.A., and Hill,D.F. (1988). *EMBO J.*, **7**, 3957–3962.
13. Sprengart, M.L., Falscher,H.P., and Fuchs,E. (1990). *Nucleic Acid Res.*, **18**, 1719–1723.
14. Faxen,M., Plumbridge,J., and Isaksson,L.A. (1991). *Nucleic Acid Res.*, **19**, 5247–5251.
15. Gold,L. (1988). *Ann Rev Biochem.*, 57, 199–233.
16. de Smit,M.H., and van Druin,J. (1990). *Biochemistry.*, **87**, 7668–7672.
17. Vellanoweth,R.L., and Rabinowitz,J.C. (1992). *Mol Micro.*, **6**, 1105–1114.
18. Rudd,K.E. (1992). In Miller,J. (ed.), A short course in bacterial genetics: A laboratory manual and handbook for Escherichia coli and related bacteria. Cold Spring Harbor Press, Cold Spring Harbor.
19. Devereux,J., Haeberli,P., and Smithies,O. (1984). *Nucleic Acid Res.*, **12**, 387–395.
20. Sharp,P.M., and Li,W-H. (1987). *Nucleic Acid Res.*, **15**, 1281–1295.
21. Sharp,P.M. (1991). *J. Mol. Evol.* **33**, 23–33.
22. Sharp,P.M., and Li,W-H. (1986). *Nucleic Acid Res.*, **14**, 7737–7749 .
23. Stormo,G.D., Schneider,T.D., and Gold,L.M. (1982). *Nucleic Acid Res.*, **10**, 2971–2996.
24. Schneider,T.D., Stormo,G.D., Gold,L., and Ehrenfeucht,A. (1986). *J Mol Biol.*, **188**, 415–431.
25. Kozak,M. (1983). *Microbiol Rev.*, **47**, 1–43.